

# Practice I

Identify a candidate gene from  
another genome and extract the  
sequence (including flanks)

# Find and upload candidate gene sequence and reference

- Look for the term ‘Amylase wheat’ on <https://www.ncbi.nlm.nih.gov/>

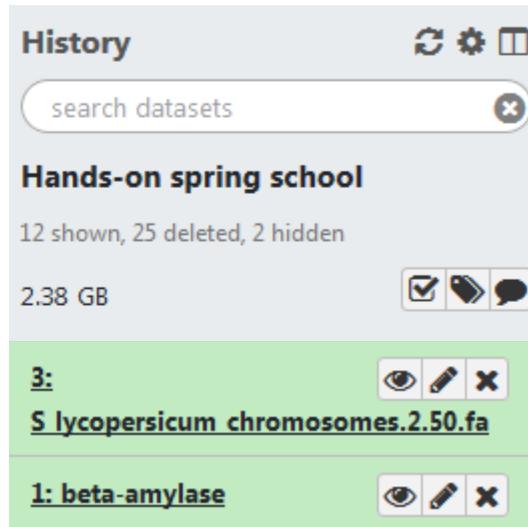
The screenshot shows the NCBI Gene search interface. The search term 'amylase wheat' is entered in the search bar. The results page displays a table of gene entries, with the entry for 'amy1' highlighted by a green box. The table columns include Name/Gene ID, Description, Location, and Aliases. The 'amy1' entry is described as 'beta-amylase [Triticum aestivum (bread wheat)]' with ID 543318.

Name/Gene ID	Description	Location	Aliases
<input type="checkbox"/> <a href="#">Ima1</a> ID: 542981	monomeric alpha-amylase inhibitor [ <i>Triticum aestivum</i> (bread wheat)]		
<input type="checkbox"/> <a href="#">0_19 alpha-AI</a> ID: 542776	0.19 alpha-amylase inhibitor [ <i>Triticum aestivum</i> (bread wheat)]		
<input type="checkbox"/> <a href="#">amy1</a> ID: 543318	beta-amylase [ <i>Triticum aestivum</i> (bread wheat)]		
<input type="checkbox"/> <a href="#">wtai-CM2</a> ID: 100682400	putative alpha-amylase inhibitor CM2 [ <i>Triticum aestivum</i> (bread wheat)]		

- Export in fasta format

# Find and upload candidate gene sequence and reference

- Get tomato reference from the ‘Galaxy workshop’ folder
- Your history should contain the following files now:



# BLAST your sequence

- Identify the location of the beta-amylase gene in the tomato genome

## Input

NCBI BLAST+ **blastn** Search nucleotide database with nucleotide query sequence(s) (Galaxy Version 0.3.1)

**Nucleotide query sequence(s)**  
1: beta-amylase  
(-query)

**Subject database/sequences**  
FASTA file from your history (see warning note below)

**Nucleotide FASTA subject file to use instead of a database**  
3: S\_lycopersicum\_chromosomes.2.50.fa  
(-subject)

**Type of BLAST**

- megablast - Traditional megablast used to find very similar (e.g., intraspecies or closely related species) sequences
- blastn - Traditional BLASTN requiring an exact match of 11, for somewhat similar sequences
- blastn-short - BLASTN program optimized for sequences shorter than 50 bases
- dc-megablast - Discontiguous megablast used to find more distant (e.g., interspecies) sequences

Don't forget to choose 'sstrand' option for the output

**Output format**  
Tabular (select which columns)  
(-outfmt)

**Standard columns**  
 Select/Unselect all

- qseqid = Query Seq-id (ID of your sequence)
- sseqid = Subject Seq-id (ID of the database hit)
- pident = Percentage of identical matches
- length = Alignment length
- mismatch = Number of mismatches
- gapopen = Number of gap openings
- qstart = Start of alignment in query
- qend = End of alignment in query
- sstart = Start of alignment in subject (database hit)
- send = End of alignment in subject (database hit)
- evalue = Expectation value (E-value)
- bitscore = Bit score

Select/Unselect all

**sstrand** = Subject Strand

frames = Query and subject frames separated by a '/'

btop = Blast traceback operations (BTOP)

qcovs = Query Coverage Per Subject

qcovhsp = Query Coverage Per HSP

# Prepare BED file from BLAST result

- Cut columns from a table ~~(cut)~~! choose correct tool!
  - 2, 9, 10, 1, 11, 13 (sname, sstart, send, qname, score, sstrand)
- Table should look like this:

1	2	3	4	5	6
SL2.50ch07	61146254	61146507	X98504.1	3.05e-24	plus
SL2.50ch07	61145356	61145679	X98504.1	2.86e-18	plus
SL2.50ch07	61146633	61146795	X98504.1	2.20e-13	plus
SL2.50ch07	61147257	61147501	X98504.1	1.14e-10	plus

- Convert ,tabular' to ,bed' format

Edit dataset attributes

Changed the type to bed.

Attributes Convert Datatypes Permissions

Change datatype Detect datatype Change datatype

New Type bed

This will change the datatype of the existing dataset but not modify its contents. Use this if Galaxy has incorrectly guessed the type of your dataset.

History search datasets

Hands-on spring school 7 shown, 9 deleted 869.54 MB

16: Cut on data 14 (edit) Metadata is being auto-detected

display in IGB View display with IGV local

1.Chrom	2.Start	3.End	4	5
SL2.50ch07	61146254	61146507	X98504.1	3.05e-24
SL2.50ch07	61145356	61145679	X98504.1	2.86e-18
SL2.50ch07	61146633	61146795	X98504.1	2.20e-13
SL2.50ch07	61147257	61147501	X98504.1	1.14e-10

# Sort the tomato coordinates

- Sort data in ascending or descending order
  - Required for bedtools merge
- Optionally, bedtools SortBED can be used

Sort data in ascending or descending order (Galaxy Version 1.1.1)

Versions Options

Sort Query

16: Cut on data 14

Number of header lines

0

These will be ignored during sort.

Column selections

1: Column selections

on column

Column: 2

in

Ascending order  
 Descending order

Flavor

Fast numeric sort (-n)  
 General numeric sort ( scientific notation -g)  
 Natural/Version sort (-V)  
 Alphabetical sort  
 Human-readable numbers (-h)  
 Random order (-R)

+ Insert Column selections

Output unique values

Yes No

Print only unique values, based on sorted key columns. See help section for details. (--unique)

Ignore case

Yes No

Sort and Join key column values regardless of upper/lower case letters. (-i)

Execute

# Merge the exons into a single feature

- bedtools MergeBED combine overlapping/nearby intervals into a single interval

**bedtools MergeBED** combine overlapping/nearby intervals into a single interval (Galaxy Version 2.27.1)

Versions Options

Sort the following BAM/bed,bedgraph,gff,vcf file  
26: Sort on data 16

Calculation based on strandedness?  
Overlaps on either strand

Maximum distance between features allowed for features to be merged  
1000

That is, overlapping and/or book-ended features are merged. (-d)

Print the header from the A file prior to results  
Yes No

(-header)

Applying operations to columns from merged intervals  
+ Insert Applying operations to columns from merged intervals

Execute

# Expand flanks to cover the whole gene

- bedtools SlopBed adjust the size of intervals

bedtools SlopBed adjust the size of intervals (Galaxy Version 2.27.1)

Versions  Options

**bed,bedgraph,gff,vcf file**  
   16: Cut on data 14

**Genome file**  
   Genome file from your history

**Genome file**  
   3: S\_lycopersicum\_chromosomes.2.50.fa (as tabular)

**Define -l and -r as a fraction of the feature's length**  
 Yes  No  
E.g. if used on a 1000bp feature, -l 0.50, will add 500 bp "upstream"

**Define -l and -r based on strand**  
 Yes  No  
If used, -l 500 for a negative-stranded feature, it will add 500 bp to the end coordinate

**Choose what you want to do**  
Increase the bed,bedgraph,gff,vcf entry by the same number base pairs in each direction.

**Number of base pairs**  
1000

**Print the header from the A file prior to results**  
 Yes  No  
(-header)

Execute

# Finally, extract the sequence

- bedtools GetFastaBed use intervals to extract sequences from a FASTA file

The screenshot shows the Galaxy web interface with the 'GetFastaBed' tool selected. The tool's header indicates it uses intervals to extract sequences from a FASTA file (Galaxy Version 2.27.0.0). A message box at the top right says 'There is a newer version of this tool available.'

The 'BED/VCF/GFF file' dropdown is set to '34: (hidden) Merged Text transformation on data 26'. The 'Choose the source for the fasta file' dropdown is set to 'History' and then 'Fasta file', which lists '3: S\_lycopersicum\_chromosomes.250.fa'. Below these, there are several configuration options:

- 'Use the 'name' column in the BED file for the FASTA headers in the output FASTA file': Yes (selected)
- '(-name)': No
- 'Report extract sequences in a tab-delimited format instead of in FASTA format': No
- '(-tab)': Yes (selected)
- 'Force strandedness': No
- 'If the feature occupies the antisense strand, the sequence will be reverse complemented. (-s)': No
- 'Treat split/spliced BAM or BED12 entries as distinct BED intervals when computing coverage.': No
- 'If set, the coverage will be calculated based the spliced intervals only. For BAM files, this inspects the CIGAR N operation to infer the blocks for computing coverage. For BED12 files, this inspects the BlockCount, BlockStarts, and BlockEnds fields (i.e., columns 10,11,12). If this option is not set, coverage will be calculated based on the interval's START-END coordinates, and would include introns in the case of RNAseq data. (-split)': No

A large green arrow points from the tool configuration on the left to the 'History' panel on the right. The 'History' panel shows the execution of '39: GetFastaBed on data 3 and data 34'. It displays 1 sequence in FASTA format:

```
>SL2.50ch07:61145356-61147501
CAGTTCAAGGAACTGAGAGAACGGGGTTGATGGGATCATGGTC
ATACTAACCGGACTGGTACAAGGAACAAAAGATGCCCTCACTTC
TAGACATAGAGGTAGGACTTGTTCTGGGGTGAGCTTAGATATC
CGAAAGTGCTTATAAGTTGGTACCCCCAACCTTATGTTTGTCAC
```

Below this, another entry '38: SlopBed on data 3 and data 29' is partially visible.

# Extract a workflow

- Use the 'Extract workflow' button in your history options
- Change name of the two inputs and the workflow itself – ready

The screenshot shows the Galaxy web interface. On the left, the 'History' sidebar is open, displaying various history lists and actions. A green arrow points from the 'Extract Workflow' button in the 'CURRENT HISTORY' section to the right-hand workflow extraction dialog.

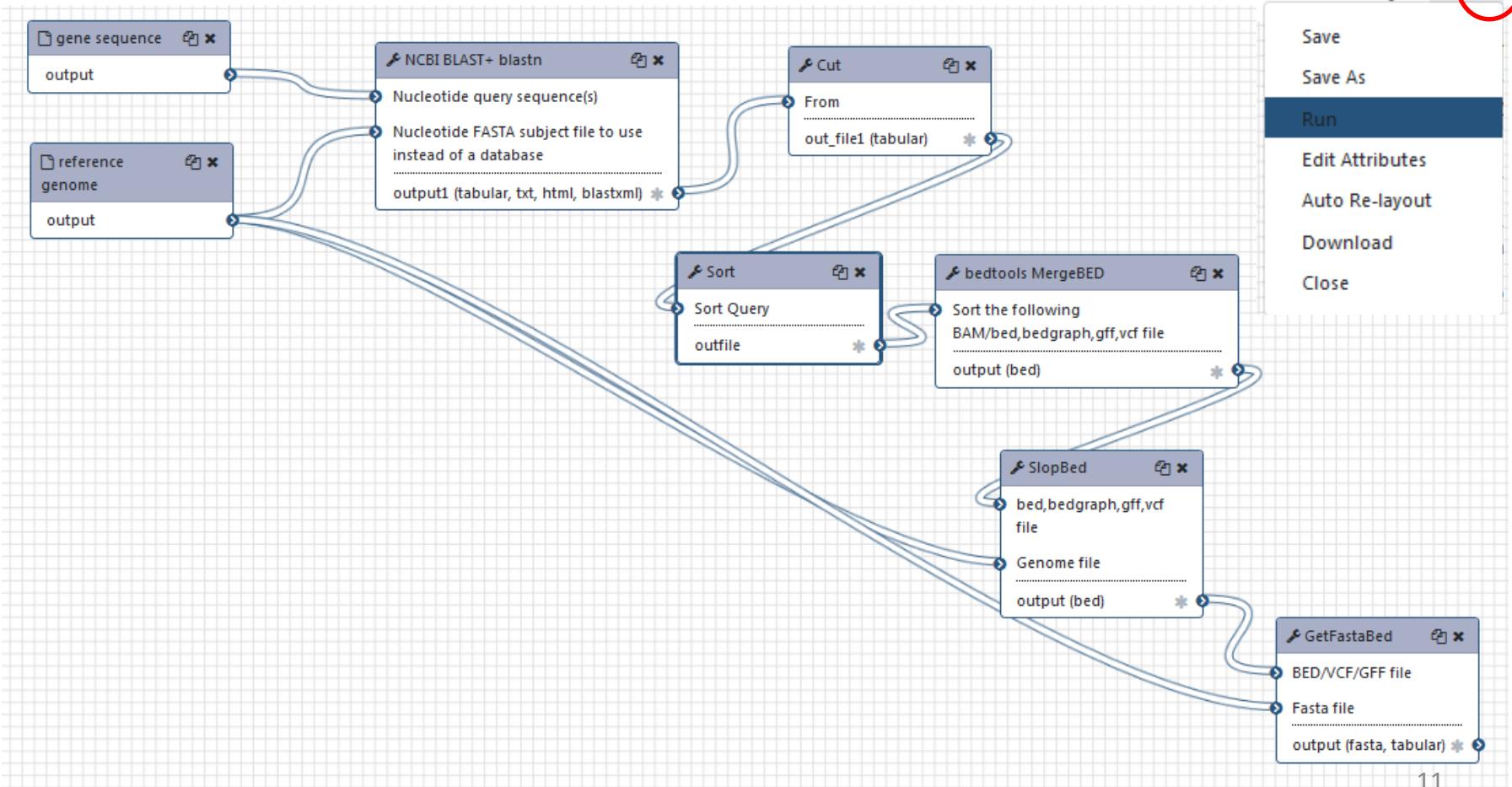
**Workflow name:** My first workflow

**Tools:**

Tool	History items created
Upload File <small>This tool cannot be used in workflows</small>	1 beta-amylase <input checked="" type="checkbox"/> Treat as input dataset Gene sequence
Upload File <small>This tool cannot be used in workflows</small>	3 S lycopersicum chromosomes.2.50.fa <input checked="" type="checkbox"/> Treat as input dataset Reference genome
NCBI BLAST+ blastn <input checked="" type="checkbox"/> Include "NCBI BLAST+ blastn" in workflow	14 blastn beta-amylase vs S lycopersicum chromosomes.2.50.fa
Cut <input checked="" type="checkbox"/> Include "Cut" in workflow	16 Cut on data 14
Sort <input checked="" type="checkbox"/> Include "Sort" in workflow	25 Sort on data 21
bedtools MergeBED <input checked="" type="checkbox"/> Include "bedtools MergeBED" in workflow	29 Merged Sort on data 16
SlopBed <input checked="" type="checkbox"/> Include "SlopBed" in workflow	38 SlopBed on data 3 and data 29
GetFastaBed <input checked="" type="checkbox"/> Include "GetFastaBed" in workflow	39 GetFastaBed on data 3 and data 34

- Some correction might be required, if so, please re-order and connect to tools
- If you have fixed the tool order, you can save and test the workflow by using the gear symbol in the right upper corner

My first workflow



- After you have started your workflow, the resulting history should like this
- First task completed



History ⚙️ ⚙️ ✎

search datasets ✖️

**My first workflow**

6 shown

4.86 KB

<a href="#">6: GetFastaBed on data 3 and data 5</a>	<span style="border: 1px solid #ccc; padding: 2px 5px;">eye</span> <span style="border: 1px solid #ccc; padding: 2px 5px;">pencil</span> <span style="border: 1px solid #ccc; padding: 2px 5px;">x</span>
<a href="#">5: SlopBed on data 3 and data 4</a>	<span style="border: 1px solid #ccc; padding: 2px 5px;">eye</span> <span style="border: 1px solid #ccc; padding: 2px 5px;">pencil</span> <span style="border: 1px solid #ccc; padding: 2px 5px;">x</span>
<a href="#">4: Merged Sort on data 2</a>	<span style="border: 1px solid #ccc; padding: 2px 5px;">eye</span> <span style="border: 1px solid #ccc; padding: 2px 5px;">pencil</span> <span style="border: 1px solid #ccc; padding: 2px 5px;">x</span>
<a href="#">3: Sort on data 2</a>	<span style="border: 1px solid #ccc; padding: 2px 5px;">eye</span> <span style="border: 1px solid #ccc; padding: 2px 5px;">pencil</span> <span style="border: 1px solid #ccc; padding: 2px 5px;">x</span>
<a href="#">2: Cut on data 1</a>	<span style="border: 1px solid #ccc; padding: 2px 5px;">eye</span> <span style="border: 1px solid #ccc; padding: 2px 5px;">pencil</span> <span style="border: 1px solid #ccc; padding: 2px 5px;">x</span>
<a href="#">1: blastn beta-amylase vs 'S lycopersicum chromosomes.2.50.fa'</a>	<span style="border: 1px solid #ccc; padding: 2px 5px;">eye</span> <span style="border: 1px solid #ccc; padding: 2px 5px;">pencil</span> <span style="border: 1px solid #ccc; padding: 2px 5px;">x</span>

# Practice II

Dataset collections and RNA-Seq  
workflow

# Import workflow

- Add the following address
  - <https://usegalaxy.eu/u/anfi/w/rna-seq/json>

---

Import Workflow

Please provide a Galaxy workflow export URL or a workflow file.

Archived Workflow URL

If the workflow is accessible via a URL, enter the URL above and click Import.

Archived Workflow File

If the workflow is in a file on your computer, choose it and then click Import.

Import a Workflow from myExperiment

[Visit myExperiment](#)

Click the link above to visit myExperiment and search for Galaxy workflows.

# References and input data

- Tomato genome sequence is already in your history (first exercise) -> please copy it to a new history
- Import tomato GFF file (link can found in course material)  
[ftp://ftp.solgenomics.net/tomato\\_genome/annotation/ITAG2.4\\_release/ITAG2.4\\_gene\\_models.gff3](ftp://ftp.solgenomics.net/tomato_genome/annotation/ITAG2.4_release/ITAG2.4_gene_models.gff3)
- Upload all fastq files from the folder  
,rnaseq\_rawdata‘

# Creating dataset collections

The screenshot shows the Galaxy software interface with two panels. The left panel displays a history of datasets titled 'GBS datasets' containing four entries: '4: CS1\_R2.mini.trim.fq', '3: CS1\_R1.mini.trim.fq', '2: ETC1\_R1.mini.trim.fq', and '1: ETC1\_R2.mini.trim.fq'. A green arrow points from step 3 to the checkbox icon in the toolbar above the list. Step 3 also includes instructions to click 'For all selected...'. The right panel shows the same history after step 3, with the first two datasets checked. A green arrow points from step 4 to the 'For all selected...' button. Step 4 also includes instructions to click 'For all selected...'. The bottom right panel shows a dropdown menu with options: Hide datasets, Unhide datasets, Delete datasets, Undelete datasets, Build Dataset List (highlighted), Build Dataset Pair, and Build List of Dataset Pairs.

1. Load your FASTQ samples in your history
2. Click „Operations on multiple datasets“ opens additional options in your history
3. Mark all files from your history that should be included in the list by clicking the small box in front of each sample.  
Go on with clicking „For all selected...“
4. This will open the following dialog:
  - Hide datasets
  - Unhide datasets
  - Delete datasets
  - Undelete datasets
  - Build Dataset List**
  - Build Dataset Pair
  - Build List of Dataset Pairs
5. Depending on the nature of sequencing data (paired or unpaired), two option are available:

Note: If your pairs are not named by „\_1“ and „\_2“ ending, you will probably see this message:

Could not automatically create any pairs from the given dataset names. You may want to choose or enter different filters and try auto-pairing again. Close this message using the X on the right to view more help.

0 unpaired forward - (4 filtered out) Choose filters Clear filters Auto-pair 1

0 unpaired reverse - (4 filtered out) 2

(no datasets were found matching the current filters)

This should be corrected by typing the naming pattern of your data.

In this example „\_R1“ and „\_R2“:

2 unpaired forward - (2 filtered out) Choose filters Clear filters Auto-pair R1

CS1\_R1.mini.trim.fq Pair these datasets

ETC1\_R1.mini.trim.fq Pair these datasets

2 unpaired reverse - (2 filtered out) R2

CS1\_R2.mini.trim.fq

ETC1\_R2.mini.trim.fq

Check the correct pairing of the files. Then click „Auto-Pair“.

The files are paired according to their name scheme. Name and create the list now.

2 pairs created: all datasets have been successfully paired

0 unpaired forward - (0 filtered out) Choose filters Clear filters R1

0 unpaired reverse - (0 filtered out) R2

2 paired Unpair all

CS1_R1.mini.trim.fq →	CS1.mini.trim	← CS1_R2.mini.trim.fq
ETC1_R1.mini.trim.fq →	ETC1.mini.trim	← ETC1_R2.mini.trim.fq

Remove file extensions from pair names?

Name: My paired list

Create list

- Your history should now show two collection, each containing 6 paired-end files
- Everything is prepared to start the workflow

**41: control**  
a list of pairs with 6 items

<u>1 GTT L0 001.fastq</u>	a pair of datasets
<u>1 GTT L0 002.fastq</u>	a pair of datasets
<u>2 GTT L0 001.fastq</u>	a pair of datasets
<u>2 GTT L0 002.fastq</u>	a pair of datasets
<u>3 GTT L0 001.fastq</u>	a pair of datasets
<u>3 GTT L0 002.fastq</u>	a pair of datasets

**40: treated**  
a list of pairs with 6 items

39: Concatenate datasets on data 38 and data 37

# Starting the workflow

- Go to ‚Workflow‘ menu and select ‚Run‘ for the RNA-Seq workflow
- Check whether the ‚GFF gene identifier‘ option (Advanced option ) is set to ‚Parent‘

Workflow: RNA-Seq for two conditions: I Sequence Mapping and Trimming and DESeq2 (imported from uploaded file)

Run workflow

Workflow Parameters

condition1  
control

condition2  
treated

History Options

Send results to a new history  
Yes No

History name  
My RNA-Seq workflow

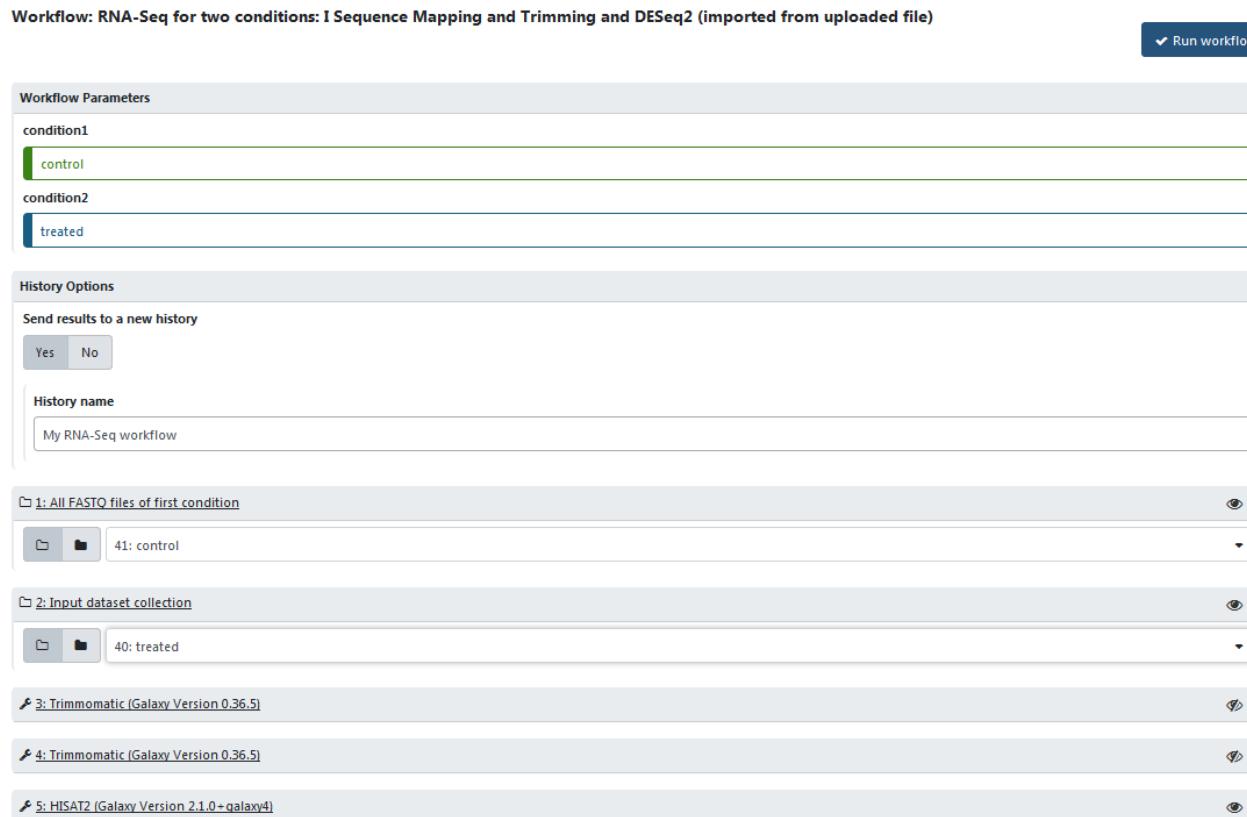
1: All FASTQ files of first condition  
41: control

2: Input dataset collection  
40: treated

3: Trimmomatic (Galaxy Version 0.36.5)

4: Trimmomatic (Galaxy Version 0.36.5)

5: HISAT2 (Galaxy Version 2.1.0+galaxy4)



- This is the final history (right)
- Please inspect the results
- Improvements:
  - Renaming and joining the samples previously
- Second task is completed



History ⚙️ ⚙️ ✎

**My RNAseq test**  
8 shown, 115 hidden  
64.87 MB  📁 💬

<a href="#">123: All raw counts</a>	<span style="border: 1px solid #ccc; padding: 2px;">eye</span> <span style="border: 1px solid #ccc; padding: 2px;">pen</span> <span style="border: 1px solid #ccc; padding: 2px;">x</span>
<a href="#">122: All FeatureCounts statistics</a>	<span style="border: 1px solid #ccc; padding: 2px;">eye</span> <span style="border: 1px solid #ccc; padding: 2px;">pen</span> <span style="border: 1px solid #ccc; padding: 2px;">x</span>
<a href="#">121: DESeq2 plots on data 116, data 115, and others</a>	<span style="border: 1px solid #ccc; padding: 2px;">eye</span> <span style="border: 1px solid #ccc; padding: 2px;">pen</span> <span style="border: 1px solid #ccc; padding: 2px;">x</span>
<a href="#">120: DESeq2 result file on data 116, data 115, and others</a>	<span style="border: 1px solid #ccc; padding: 2px;">eye</span> <span style="border: 1px solid #ccc; padding: 2px;">pen</span> <span style="border: 1px solid #ccc; padding: 2px;">x</span>
<a href="#">88: featureCount on treated</a> a list with 6 items	<span style="border: 1px solid #ccc; padding: 2px;">x</span>
<a href="#">67: featureCounts on control</a> a list with 6 items	<span style="border: 1px solid #ccc; padding: 2px;">x</span>
<a href="#">60: HISAT2 on treated</a> a list with 6 items	<span style="border: 1px solid #ccc; padding: 2px;">x</span>
<a href="#">53: HISAT2 on control</a> a list with 6 items	<span style="border: 1px solid #ccc; padding: 2px;">x</span>

# Bonus

- Please edit the workflow to include
  - Fastqc to check read quality
  - a clipping tool (e.g. Trimmomatic)

before the HISAT2 mapping step